Audio Captioning モデルの発達的カリキュラム学習 *

◎稲垣 賢斗, 高道 慎之介 (慶大)

1　はじめに

オーディオキャプショニング [1] は，入力音に対して，その内容を説明するキャプションを生成するクロスモーダルな機械学習タスクである．入力されるオーディオは，環境音や生活音などを含むものであり，モデルはそれらの前景・背景の出来事を自然言語で記述するように学習される．

こうした言語とオーディオなどのクロスモーダルな接続は，より汎用的な機械学習モデルの発展において重要な役割を果たす．近年では，大規模言語モデル（large language model; LLM）といった自然言語処理に特化した高性能なモデルが登場しているが，これらのモデルは依然として人間の能力に及ばない点がある．特に複数のモーダル間の統合的な処理能力において大きな差があり，マルチモーダル LLM [2] の研究は盛んに行われているが，人間のように複数のモーダルを自然に扱う能力にはまだ到達していない．

機械学習と人間の違いに着目することは，今後のモデル設計における有益な示唆を与えると考えられる．例えば，人間の学習過程には，無作為な情報提示ではなく，意味のある順序で段階的に情報が提示されることで効率が高まるという特徴がある．こうした学習戦略は機械学習にも応用可能であるとされ，カリキュラム学習（curriculum learning）[3] として注目を集めている．特に言語獲得においては，単語とモノの対応のような単純な結びつきから始まり，徐々に文法構造や抽象概念の理解へと発展するという発達的な段階性が見られる．

本研究では，こうした人間の学習戦略をオーディオキャプショニングに応用した，発達的カリキュラム学習を提案する．発達的カリキュラム学習では訓練時，Fig. 1 に示すように，人間の言語発達に倣って段階的にキャプションの内容を調整する．初期はキーワード中心の簡素な文から始め，後期には元の完全な文に近づけるといったカリキュラムを実装することで，モデルの性能や学習過程にどのような変化が生じるかを調査する．

2　関連研究

2.1　日本語の言語発達

人間の言語獲得について論じるにあたっては，言語によって発達の特性が異なる点に留意する必要がある．本研究では日本語を対象とし，日本語における幼児の言語発達の特徴に基づいて議論を行う．

幼児は言語発達の始まりとして喃語（なんご）を発する．これは，音声学的には言語と無関係とされているが，初語（しょご）との間に明確な連続性があることがわかってきており，環境に対してフィードバック的に発声されるという報告もある [4]．その後，生後

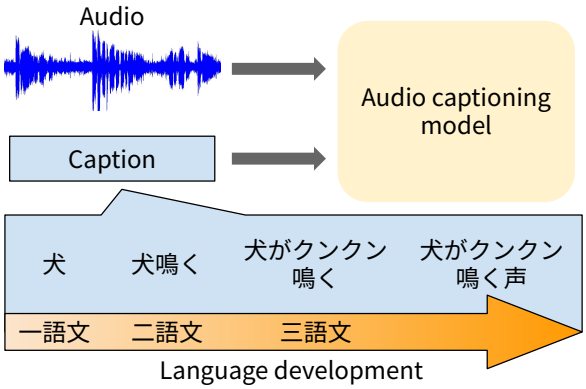

Fig. 1　オーディオキャプショニングにおける発達的カリキュラム学習の概要．

1 年前後から初語を発し始め，そこから急速に言語能力を発達させていく．語彙や文法の発達には一定の順序性が見られ，例えば初期語彙には名詞が多く含まれ，文法の発達が進むにつれて動詞や形容詞などの使用頻度が高まることが知られている [5]．また，典型的には 1 歳半から 2 歳ごろに二語文（二語の組み合わせ）を話すようになり，2 歳から 3 歳にかけては，三語以上の文や助詞・動詞の活用を含む発話が現れるようになる [6]．本研究では，このように文の長さや文法的複雑性が段階的に上昇する幼児の言語発達に倣い，発達的カリキュラム学習の設計に応用する．

2.2　BabyLM

BabyLM [7] は，「人間の言語学習から得られる洞察は，言語モデルの改良にどのように活かせるか？」という，本研究と近い理念のもと開催されたコンペティションである．ここでは，人間が母語を完全に扱えるようになるとされる 13 歳ごろまでに触れる言葉と同程度の量のデータ（1 億語未満）を用いて言語モデルを訓練するという課題が与えられている．BabyLM は，言語発達といった認知科学的知見が高効率な言語モデルの構築にどのように活かせるかを問うとともに，言語モデルのアーキテクチャが認知科学の研究にどのようなインスピレーションを与えうるかという双方向的な関心を持って設計されている．

本コンペティションでは，カリキュラム学習によって性能向上を目指す手法も多数提案されている．例えば，マスク言語モデル（masked langage model）において，標準的な子どもの言語発達スケジュールに基づいてカリキュラムを設計し，マスク対象となる語彙を段階的に増やしていく手法 [8] では，評価スコアの改善と学習速度の向上が報告されている．しかし，コンペティション全体としては，現時点ではカリキュラム学習によって特別顕著な成果が得られていると

*Developmental Curriculum Learning for Audio Captioning Models, by Kento Inagaki, Shinnosuke Takamichi (Keio University).

は言い難く，今後のさらなる検討が求められている．

2.3 画像キャプショニングのカリキュラム学習

画像キャプショニング（image captioning）は画像を自然言語で説明するタスクである．カリキュラム学習を活用した研究事例は多数見られ，難易度によりデータセットを並び替える手法や [9, 10]，単語，フレーズ，文の順に予測の粒度を上げていく手法 [11]，本研究と同様に子供の言語習得から着想を得た手法 [12] などが研究されている．

2.4 ストップワード削除カリキュラム

オーディオキャプショニングモデルにおけるカリキュラム学習の手法として，ストップワード削除カリキュラム学習 [13] がある．この手法ではキャプションからストップワード（for, do, the など情報量が少なく，頻繁に使われる単語）を確率的に削除することで，主要な意味を伴うキーワードの比率を高め，キャプション予測の難易度を低減する．カリキュラム学習の枠組みとしては，学習初期のエポックでキーワードの比率が高い簡易なキャプションを用い，学習が進むにつれてその比率を徐々に戻し，最終的には元の完全なキャプションに近づけていく．このような段階的に難易度を上げる手法で、モデルの性能向上が報告されている．エポック (epoch) ごとのストップワード削除確率 P は次式で定義される：

$$P = \min\left(e^{-\alpha * \text{epoch}}, 0.95\right) \tag{1}$$

3 提案手法

本研究では，幼児の言語発達に倣って，段階的に文長・文法複雑性が上昇するキャプションのデータセットによる発達的カリキュラム学習を提案する．4 段階のカリキュラムを設け，各段階ごとの幼児発話モデリングに応じて元キャプションを加工する．Table 1 にキャプションの例を示す．各段階のモデリングとキャプション加工のルールを次に示す．

キャプションには複数のイベントが含まれることがある．多くの場合，句読点によって複数のイベントの記述が区切られるため，キャプション中の句読点で分割した部分ごとに以下の加工を行う．

3.1 第一段階

幼児の発話の最初期にみられる一語文をモデリングする．初期語彙に名詞が多く含まれるという報告 [5] から，キャプションから名詞を一つ抽出し，それを第一段階のキャプションとする．

3.2 第二段階

二語文をモデリングする．二語文は通常（名詞＋動詞）や（形容詞＋名詞）で構成さる．第一段階で抜き出した名詞と，加えて動詞や形容詞を一つ抽出し，それらを第二段階のキャプションとする．

3.3 第三，四段階

第三段階では，文法的な構成が始まる段階をモデリングする．第二段階の二語文に一語加えた三語文を構成し，それらの主要な三語に加えて助詞と助動詞を追加する．第四段階では完全な文として元キャプションを用いる．

4 実験的評価

本研究における提案手法，発達的カリキュラム学習がオーディオキャプショニングにおいて有効であるか，実験する．

4.1 モデル

実験に使用するオーディオキャプショニングモデルとして，本研究では CNext-trans [14] を採用した．このモデルは，オーディオキャプショニングのコンペティションである DCASE 2024 Task 6 [15] においてベースラインとして用いられたものである．モデルは，事前学習済みの畳み込み型オーディオエンコーダである ConvNeXt [16] と，それぞれ 8 つのヘッドアテンション機構を持つ 6 層の Transformer [17] デコーダで構成される．オーディオキャプショニングタスクにおいて優れた性能を示しているモデルの多くは，デコーダ部に事前学習済みの言語モデルを用いているが，本研究は人間の言語獲得を模倣した学習を目的としているため，言語データによる事前学習が行われていないモデルを採用した．

4.2 比較手法

比較手法は，学習全体を通して同じデータセットを用いるカリキュラム無し学習，従来手法のストップワード削除カリキュラム学習，および提案手法の発達的カリキュラム学習の 3 つである．使用したデータセットは，約 10 秒の環境音と日本語キャプションのペアから構成される Multi-lingual AudioCaps [18] である．データ数は，train：45,205，validation：440，test：883 であり，validation と test は 1 つのオーディオに対して 5 つのキャプションが付与されている．従来手法および提案手法では，学習データセットのキャプションを加工することで，各々のカリキュラムを構築する．キャプションの加工には，形態素解析ツール Janome[1]を用いた．従来手法において，ストップワード削除確率の変数は $\alpha = 0.05$ に設定し，日本語におけるストップワードの選定は [19] を参照した．提案手法のカリキュラムは，20 エポックまで第一段階，40 エポックまで第二段階，60 エポックまで第三段階，その後を第四段階とした．3 手法共通のパラメータとして，エポック数は 100，学習率は 5×10^{-4} に設定した．

4.3 評価指標

生成キャプションの評価指標として，BLEU [20]，ROUGE-L [21]，BERTScore [22]，および CLAIR-A [23] を用いた．BLEU は，生成キャプションと参照キャプションとの n-gram の重なり具合を測定する指標であり，ROUGE-L は最長共通部分列（longest common subsequence; LCS）に基づいてスコアを算出する．BERTScore は，事前学習済みの言語モデ

[1] https://github.com/mocobeta/janome

Table 1　発達的キャプションの例

第一段階	第二段階	第三段階	第四段階（元キャプション）
車	車エンジン	車のエンジンの作動	複数の車のエンジンの作動音
大人、猫	大人話し、猫鳴く	大人の女性が話し、猫がニャーと鳴く	大人の女性が話している声、猫がニャーと鳴く声

Table 2　実験結果

カリキュラム	BLUE	ROUGE-L	BERTScore	CLAIR-A
無し	0.313	0.501	0.834	0.648
従来手法	0.312	0.502	0.835	**0.650**
提案手法	**0.322**	**0.508**	**0.836**	0.641

Table 3　最終エポックにおける生成キャプションの例

正解キャプション	カリキュラム無し	提案手法
女性の歌声。その後で、咳。それに続いて、鳥たちがチーチー鳴く声	男性が話している声。その後に、赤ちゃんの泣き声	鳥たちがチューチュー鳴く声
女性の話している声。それに続いて、磁器の皿がガチャンと鳴る音、食べ物と油がシューシューいう音	女性が話している声、同時に食べ物を炒める音	女性が話している声、同時に食べ物を炒められる音

ル BERT から得られるベクトル表現を用いて，生成キャプションと参照キャプションの意味的類似度を評価する．CLAIR-A は，LLM に生成キャプションの質を点数で評価させたものである．LLM は ChatGPT-4o [2]を使用した．日本語対応のため，英文プロンプトを ChatGPT-4o を用いて日本語翻訳した．

4.4　結果

最終エポックにおけるモデルの評価結果を Table 2 に示す．BLEU，ROUGE-L，BERTScore において提案手法が最も高い値となった一方で，CLAIR-A においては最も低い値となった．Table 3 上の例のように，カリキュラム無しよりも提案手法の方が正しくイベントを記述できているケースが多くみられた一方で，下の例ような文法の間違いは提案手法の方で多くみられた．test set 884 個の中で文法エラーが含まれるキャプションを数えたところ，カリキュラム無しでは 11 個だったのに対して，提案手法では 42 個であった．CLAIR-A では LLM に対し文法エラーも点数に反映するように指示しているため，それが主に影響して低い結果になったと考えられる．提案手法では，初期段階で音響イベントを記述する上で主要な単語を中心に学習させることで，オーディオと単語の対応は改善したものの，正しい文法で学習される段階が減るため，文法のエラーが増えてしまったと考えられる．

4.5　学習過程における考察

Table 4 は各手法の 10 epoch 目における出力である．カリキュラム無しでは「男性が話している声」が頻出し，音の特徴を捉えていないキャプションになっているが，提案手法では正解キャプションに含まれる単語を出力をしている．教師あり学習では出力の誤差が最小化されるように学習されるため，初期段階では教師データに頻出するパターンが一様に出力される傾向がある．そこでカリキュラム学習では，提示する教師データを段階的に変化させることで，学習の道筋を誘導することができる．ただし，Table 4 上の例では，カリキュラム無しの学習において，10 epoch 目の出力は「男性が話している声」であったが，最終エポックでは「女性が話している声、同時に食べ物を炒める音」と，音を正しく表すキャプションを出力できている．これは，提案手法の最終エポックと同様の出力であり，学習の過程は異なるが，最終的には同様な出力となっている．学習の途中過程が，最終的なモデルの内部構造にどのように影響を与えているのかを調査することはブラックボックス的な機械学習モデルにおいては難しく，課題として残る点である．

Table 4　10 epoch 目における生成キャプションの例

正解キャプション	カリキュラム無し	提案手法
食べ物が炒められる音、女性の話している声	男性が話している声	女性、水
エンジンがゴロゴロ鳴る大きな音。その後で、エアホーンが鳴る音が三度	男性が話している声	エンジン、車両
子供の泣き声、男性と女性の話している声。それに続いて、車のドアが開く音。その後で、閉まる音	男性が話している声、それに続い始める	子供、女性

5　まとめ

本研究では，オーディオキャプショニングモデルにおいて，人間の言語獲得過程を模倣した発達的カリキュラム学習を提案し，その有効性を示した．具体的には，幼児の日本語獲得の過程で現れる一語発話，二語発話などの段階をモデリングしてキャプションを加工し，段階的に分長・文法複雑性が増加するデータセットでのカリキュラム学習を行った．提案手法は，単語・意味の一致度を示す評価指標において，カリキュラム無し・従来手法に比べて高い値を示し，オーディオとキャプションの意味的対応の改善がみられた一方で，文法のエラーが増加した．この結果は，機械学習モデルにおいて人間の学習戦略の応用が有効である可能性を示唆する．

謝辞:本研究は，JST 創発的研究支援事業 JPMJFR226V，JSPS 科研費 23K24895 の支援を受けて実施した．

[2] https://openai.com/ja-JP/index/hello-gpt-4o/

参考文献

[1] X. Mei et al., "Automated audio captioning: an overview of recent progress and new challenges," *EURASIP J. Audio Speech Music Process.*, vol. 2022, no. 1, Oct. 2022.

[2] Z. Liang et al., "A survey of multimodel large language models," in *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, ser. CAICE '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 405–409.

[3] Y. Bengio et al., "Curriculum learning," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09. New York, NY, USA: Association for Computing Machinery, 2009, p. 41–48.

[4] C. Laing, E. Bergelson, "From babble to words: Infants' early productions match words and objects in their environment," *Cognitive Psychology*, vol. 122, p. 101308, 2020.

[5] 小椋たみ子, "日本の子どもの初期の語彙発達," **言語研究**, vol. 132, pp. 29–53, 2007.

[6] 岡本夏木, **子どもとことば**. 岩波新書, 1982.

[7] M. Y. Hu et al., "Findings of the second BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora," in *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, M. Y. Hu et al., Eds. Miami, FL, USA: Association for Computational Linguistics, Nov. 2024, pp. 1–21.

[8] E. Lucas et al., "Using curriculum masking based on child language development to train a large language model with limited training data," in *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, M. Y. Hu et al., Eds. Miami, FL, USA: Association for Computational Linguistics, Nov. 2024, pp. 221–228.

[9] H. Zhang et al., "Cross-modal similarity-based curriculum learning for image captioning," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg et al., Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 7599–7606.

[10] M. Alsharid et al., "A course-focused dual curriculum for image captioning," in *IEEE International Symposium on Biomedical Imaging*, Dec. 2021, p. 716–720.

[11] T. Liu et al., "Multi-stage pre-training over simplified multimodal pre-training models," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, C. Zong et al., Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 2556–2565.

[12] H. Ayyubi et al., "Learning from children: Improving image-caption pretraining via curriculum," in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers et al., Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 13 378–13 386.

[13] A. Koh et al., "Automated audio captioning with epochal difficult captions for curriculum learning," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 1058–1063.

[14] E. Labbé et al., "Conette: An efficient audio captioning system leveraging multiple datasets with task embedding," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 32, p. 3785–3794, Jul. 2024.

[15] D. Epshtein et al., "Dcase 2024 task6: Automated audio captioning using contrastive learning," DCASE2024 Challenge, Tech. Rep. 34, May 2024, ranked 10/12 in the DCASE2024 Challenge Task 6 with FENSE score of 0.514.

[16] Z. Liu et al., "A convnet for the 2020s," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11 966–11 976.

[17] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems, 30*, 2017.

[18] 岡. 悠希 et al., "環境音に対する日本語自由記述文コーパスとベンチマーク分析," in **言語処理学会 全国大会**, 2024.

[19] 國. 久嗣 et al., "内容推測に適したキーワード抽出のための日本語ストップワード," **日本感性工学会論文誌**, vol. 12, no. 4, pp. 511–518, 2013.

[20] K. Papineni et al., "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ser. ACL '02. USA: Association for Computational Linguistics, 2002, p. 311–318.

[21] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81.

[22] T. Zhang et al., "Bertscore: Evaluating text generation with BERT," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[23] T.-H. Wu et al., "Clair-a: Leveraging large language models to judge audio captions," 2024. [Online]. Available: https://arxiv.org/abs/2409.12962